MiniCPM-V: A GPT-4V Level MLLM on Your Phone

Year: 2024 | Citations: 824 | Authors: Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui

---

Abstract

The recent surge of Multimodal Large Language Models (MLLMs) has fundamentally reshaped the landscape of AI research and industry, shedding light on a promising path toward the next AI milestone. However, significant challenges remain preventing MLLMs from being practical in real-world applications. The most notable challenge comes from the huge cost of running an MLLM with a massive number of parameters and extensive computation. As a result, most MLLMs need to be deployed on high-performing cloud servers, which greatly limits their application scopes such as mobile, offline, energy-sensitive, and privacy-protective scenarios. In this work, we present MiniCPM-V, a series of efficient MLLMs deployable on end-side devices. By integrating the latest MLLM techniques in architecture, pretraining and alignment, the latest MiniCPM-Llama3-V 2.5 has several notable features: (1) Strong performance, outperforming GPT-4V-1106, Gemini Pro and Claude 3 on OpenCompass, a comprehensive evaluation over 11 popular benchmarks, (2) strong OCR capability and 1.8M pixel high-resolution image perception at any aspect ratio, (3) trustworthy behavior with low hallucination rates, (4) multilingual support for 30+ languages, and (5) efficient deployment on mobile phones. More importantly, MiniCPM-V can be viewed as a representative example of a promising trend: The model sizes for achieving usable (e.g., GPT-4V) level performance are rapidly decreasing, along with the fast growth of end-side computation capacity. This jointly shows that GPT-4V level MLLMs deployed on end devices are becoming increasingly possible, unlocking a wider spectrum of real-world AI applications in the near future.